Experimental Study on Finding Audio-visual Association by Maximizing Mutual Information

Kamal Dheeriya Badgotia
Roll no. 213EC6257

Department of Electronics and Communication Engineering
National Institute of Technology, Rourkela
Rourkela, Odisha, India
June, 2015

Experimental Study on Finding Audio-visual Association by Maximizing Mutual Information

Thesis submitted in partial fulfillment of the requirements for the degree of

Master of Technology

in

Signal and Image Processing

by

Kamal Dheeriya Badgotia

Roll no. 213EC6257

under the supervision of

Dr. Lakshi Prosad Roy

Department of Electronics and Communication Engineering

National Institute of Technology, Rourkela

Rourkela, Odisha, India

June, 2015

dedicated to my guide and my parents...

National Institute of Technology
Rourkela

DECLARATION

I declare that

1. The work contained in the thesis has been done by myself under the supervision of my supervisor.
2. The work has not been submitted to any other Institute for any degree or diploma.
3. I have followed the guidelines provided by the Institute in writing the thesis.
4. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.
5. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Kamal Dheeriya Badgotia

National Institute of Technology
Rourkela

CERTIFICATE

This is to certify that the work in the thesis entitled **"Experimental Study on Finding Audio-visual Association by Maximizing Mutual Information"** submitted by ***Kamal Dheeriya Badgotia*** is a record of research work carried out by him under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of Master of Technology in Electronics and Communication Engineering (Signal and Image Processing), National Institute of Technology, Rourkela. Neither this thesis nor any part of it, to the best of my knowledge, has been submitted for any degree or academic award elsewhere.

Dr. Lakshi Prosad Roy
Assistant Professor
Department of ECE
National Institute of Technology
Rourkela

Acknowledgment

This work is one of the most important achievements of my career. Completion of my project would not have been possible without the help of many people, who have constantly helped me with their full support for which I am highly thankful to them.

First of all, I would like to express my gratitude to my supervisor **Dr. Lakshi Prosad Roy**, who has been the guiding force behind this work. I want to thank him for giving me the opportunity to work under him. He is not only a good Professor with deep vision but also a very kind person. I consider it my good fortune to have got an opportunity to work with such a wonderful person.

I am also very obliged to **Prof. K.K. Mahapatra**, HOD, Department of Electronics and Communication Engineering for creating an environment of study and research. I am also thankful to Prof. S. Meher, Dr. S. Ari, Dr. A.K. Sahoo, Dr. D.P. Acharya, Mr. A.K Swain and Mr. S. Maiti for helping me how to learn. They have been great sources of inspiration.

I would like to thank all faculty members and staff of the ECE Department for their sympathetic cooperation. I would also like to make a special mention of the selfless support and guidance I received from PhD Scholar Mr. Bibhuti Bhusan Pradhan, Mr. Dheeren Ku Mahapatra and my friends during my project work.

When I look back at my accomplishments in life, I can see a clear trace of my family's concerns and devotion everywhere. My dearest mother, whom I owe everything I have achieved and whatever I have become; my beloved father, who always believed in me and inspired me to dream big even at the toughest moments of my life; and sisters; who were always my silent support during all the hardships of this endeavor and beyond.

Kamal Dheeriya Badgotia

Abstract

In this thesis, the fusion of audio and visual signal are done by considering one camera and one microphone. Although there are many approaches for audio-visual fusion, but still it is an open problem.

Here, localizing and recognizing the speaker in a video by maximizing the mutual information occuring from audiovisual scene. Experiment is done using a method in which the arrival of audio and visual signals from a unique source are detected using a signal-level fusion technique. The information theoretic measure of audio-visual correspondence using probabilistic multi-modal generation statistical implicit model given in [1] is used here. In the used method non-parametric statistical density is able to give better result for characterizing the mutual information between signals from different domains by following some constraint for which entropy is maximized. By maximizing the above mutual information between different pairs of signals, it is possible to identify which person is speaking a given utterance, and we can verify whether the audio is associated with that visual seen or not.

In doing so, no assumption is made about user appearance or speech of person. Further more this method doesn't required any training of corpuses. In this thesis experimental results are demonstrated using CUAVE database.

Contents

List of Figures

List of Acronyms

Acronym	Description
AVSR	Audio Visual Speaker Recognition
RV	Random Variable
MI	Mutual Information
ITL	Information Theoretic Learning
MFCC	Mel Frequency Cepstral Coefficient
LPC	Linear Predictive coding
PDF	Probability Density Function
MSE	Mean Square Error
ISE	Integral Square Error

Chapter 1

Introduction

Background

Literature Reviewed

Motivation

Objective

Thesis Organization

Chapter 1

Introduction

1.1 Background

The perception that we have about the world is influenced by elements of diverse nature. Human routinely combines the information coming from different sensory modalities in order to support accurate perception. For example, with seeing and smell of the food we can guess the taste of food without actually eating it. The correspondence between produced sound and lip movement can be exploited by any listener.

The McGruk effect shows, that when human observes conflicting audio and visual stimuli, the received sound may not exist in either modality. This effect forms the basis for modeling audio and visual speech into the field of joint processing of audio-visual signal. Thus, we can say that speech is inherently bimodal, based on the clues from audio and visual signals for perception. Use of audio and visual signal adds additional modality for speaker recognition.

From this observation, researchers have been trying to combine information from different research domain. The domain in which audio only recognition is practically useful are reservations, ticket booking, traffic information, radio, fm and database access. However, these conversational speech systems works for a single user and require tethered interaction owing to which user must have to speak in telephone headset or attached with a microphone. This limits the performance of a dialogues system, since there must be circumstances where

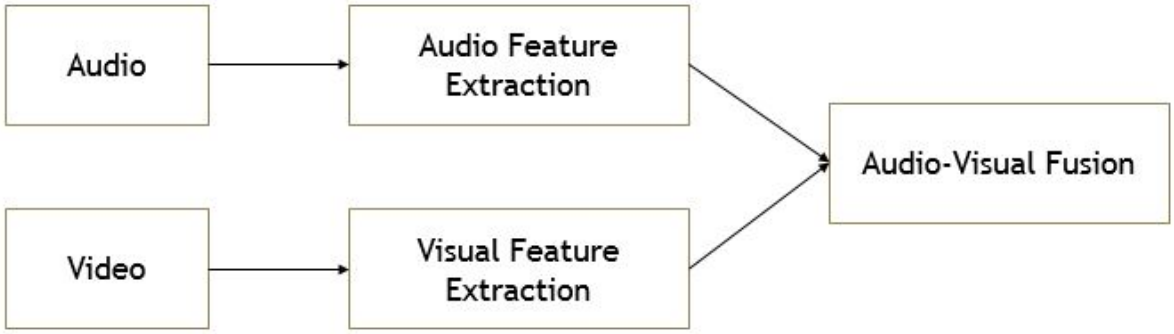


Figure 1.1: Generalized structure of audio-visual fusion

users might expect to freely communicate with a device e.g. perceptual user interfaces. Where user wish to directly command the system. So we need to localize the speaker, who can give command to the system, and it can verify that both modality belongs to the same event.

Having a single modality, opportunely associating speech from multiple unknown speaker is quite arduous. However, if other modalities are available they can often provide reliable and detailed information. In particular, visual information can be valuable for deciding whether an individual user is speaking a particular utterance. Therefore using audio visual correspondence, having a set of audio and visual signal, helps in deciding which audio visual pair signal is consistent and could have come from a single speaker.

1.2 Literature Reviewed

Initial work for fusion of audio and video using mutual information is done by Hershey and Movell [2], which shows correspondence between audio and video measured by finding the Pearson correlation coefficient between pixel intensity and average acoustic energy. They have initially assumed that the densities are Gaussian later on they have generalized it and fusing correlation coefficient in the form of mutual information, that is extracting the mutual information between each pixel and energy of audio.

J Fisher and T Darrell [1], [3] has developed another method for audio visual correspondence using mutual information and information theoretic learning. They have projected high dimensional visual and audio signal to low dimensional subspace by projection to output subspace. Then by information theoretic learning and Non-parametric density estimator, show how it is possible to capture complex dependencies between audio and video by their method.
Slaney and Covell [4] uses canonical correlation for temporal alignment between audio and video. They did not gave a way for finding whether the two signal came from the same person although their method can be adopted to do so.
Nock [5] considered two mutual information one HMM based approach for judging speech and face consistency. Most of them uses cepstral coefficient for speech signal and optical flow followed by taking discrete cosine transform of optical flow for representation features of video.

1.3 Motivation

Measuring syncroncy of audio and visual scene helps in many application which are dealing with audio-visual sequence such as Sound source localization, Teleconferencing, Speech separation by audio-visual correspondence, liveness test, Synchronizing video using audio-visual signal, Speaker localization and verification.

It is found that the main challenge in combined audio-visual processing is to model the joint density of audio and video such that it will give efficient representation of it for speaker recognition. Various way of correspondence measure of audio and visual data are found in existing literature.

Recently research that have been carried out for joint processing of audio and visual signal show that, there is equal importance of audio clues and visual clues for efficient speaker recognition, which form the basis for joint processing

of audio and visual signal. It indicates that there must exist a relation between audio and visual signal. It is also found that if a relation between audio and visual data exist then it can be associated by collecting separate data for audio and video. However in an audio-visual scene both the data are to be collected in coherent time interval. Therefore, it is required to regressly examine what happens if audio and video data do not belongs to the same event.

1.4 Objective

The primary goal of thesis work is to show whether a separately record audio and video belongs to the same event, and to give most reliable and detailed information regarding that event. Following are the objective to achieve the goal.
1. Representation of the features of audio and video, such that they can be efficiently fused.
2. Estimation of joint density of audio and video from an audio-visual scene.
3. Evaluation of audio-visual correspondence in experimental data set.
4. Examination of effectiveness in speaker verification if.
(a) Audio and video belongs to the same event.
(b) Either audio and video is not of an event.

1.5 Thesis Outline

Chapter 2 Gives an overview for finding features which gives relevent information for performing efficient recognition from audio and video, later on the methods for audio video correspondence and estimation of joint density from an audio-visual sequence.
Chapter 3 Provide in-depth review of hypothesis test for data association and how that hypothesis is related for finding mutual information, later probabilistic statistical implicit model for modeling joint density with a common audiovisual

source and background interference in each modality is show to find a equivalent Markov Chain Model from which it is shown that how the measurement of audio and video are related to the common source using data processing inequality.

Chapter 4 Gives the criteria for feature extraction, and mapping the random variable from input subspace to output subspace and show how the learning is done in the output space. The approach is driven by minimizing the integral square error between uniform density to the density of projection. later the derivation of Information Theoretic Learning algorithm is derived for adaptation of the feature vectors.

Chapter 5 Signal level fusion of audio and video to produce a result that give most reliable and detailed information. Joint processing of audio and video by signal level fusion by projecting the data from input subspace to output subspace. Adaptation of feature vector by ITL derived from mutual information is applied and result are shown using CUAVE audio-visual dataset.

Chapter 6 In this chapter experimental results are shown using CUAVE audio-visual dataset.

Chapter 2

Feature Extraction and Measure of Correspondence of Audio-visual Signal

Introduction

Audio Feature Extraction

Visual Feature Extaction

Methods for Audio-Visual Correspondence

Chapter 2

Feature Extraction and Measure of Correspondence of Audio-visual Signal

2.1 Introduction

Feature extraction is an important aspect for Audio-Visual Speaker Recognition(ASVR). For efficent recognition feature are to be carefully chosen such that it helps in extracting relevent information for performing efficient recognition from speaker's audio and image. Researcher have most tried to use feature from audio-visual data than by using raw audio and video. The McGruk effect shows, that when human observes conflicting audio and visual stimuli, the received sound may not exist in either modality. This effect forms the basis for modeling audio and visual speech into the field of joint processing of audio-visual signal. Thus audio and visual speech clues will give equal relevent information.

2.2 Audio Feature Extraction

2.2.1 Energy of Audio Signal

The average acoustic energy of audio signal is a particular audio frame is first used by Hershey and Movellan [2] as a audio feature, they have tried to find out the mutual information between audio energy and pixel intensity. The energy

of audio signal is calculated as

$$Energy = \frac{1}{N}\sum_{i=1}^{N} x(n)^2 \tag{2.1}$$

where, N is the no. of audio sample in a particular frame, $x(n)$ is the amplitude of audio signal samples. These are most used most to distinguish voiced speech and unvoiced speech. log energy and root mean square of audio amplitude was also proposed in [6] [7].

2.2.2 Mel-Frequency Cepstral Coefficients(MFCC)

MFCC was proposed in speech and audio visual speaker recognition methods [5], [8], [4], [9] because it gives the compact representation of the spectral characteristic of persons voice.

2.2.3 Linear Predictive Coding(LPC)

It is assumed that the characteristics of vocal tract remain fixed for a duration of $10ms$. The speech that we perceive, as given in [10], [11] can ne viewed as the convolution of the vocal tract and input excitation from the source.

$$S(n) = E(n) * V(n) \tag{2.2}$$

vocal tract vibrate at a particular frequency known as resonant frequency, will allow vocal tract to view as a filter, therefore in frequency domain

$$S(Z) = \frac{G}{1 + \sum_{i=1}^{q} a_i z^{-1}} \tag{2.3}$$

these, coefficient are called linear predictive coefficients[] because present sample can be predicted by the knowledge of the past q samples.

$$S(n) = \sum_{i=1}^{q} a_i S(n-i) + G \times u(n) \tag{2.4}$$

where, G is the gain and $S(n)$ is the corresponding speech samples.

2.2.4 Periodogram

J Fisher and T Darell [1] used periodogram for parametrization of speech. Periodogram is defined as the square magnitude of the FFT bins. Periodogram are based on the definition of power spectral density.

$$\begin{aligned} P_{a,N}(\omega) &= \frac{1}{N}|DTFT(a_w)|^2 \\ &= \frac{1}{N}\left|\sum_{n=0}^{N-1} a_w(n)\, e^{-j\omega n}\right|^2 \end{aligned} \tag{2.5}$$

where, $a_w(n)$ is the windowed segment of sample, $w(n)$ is the window function, N is the no. of samples.

$$a_w(n) = a(n)\, w(n) \tag{2.6}$$

2.3 Visual Feature Extraction

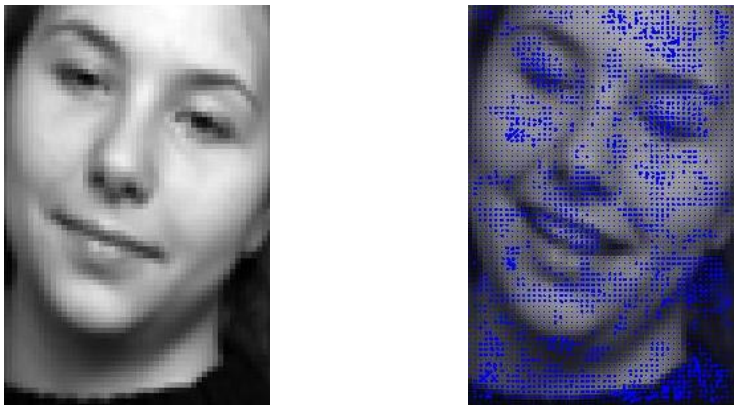
Most of the previous work of researcher tried to find out the region of interest i.e. mouth, and various other feature that we are discussing now.

2.3.1 Raw Pixels

The intensity of the gray scale image was used by [2], which is equivalent to raw audio energy. In there method, they have found per pixel correlation relative to average acoustic energy.

2.3.2 Optical Flow

Apparent motion between the object, surface and edges in any visual scene, can be easily estimated by Optical Flow which was caused by relative motion between successive visual frames, in other word it can also be said as as relative motion between an observer(an eye or a camera) and the scene. velocity of that pixel movement can be calculated by Horn and Shrunk method given in [12].



(a) (b)

Figure 2.1: (a) image from a video sequence. (b) optical flow between two consecutive frame, arrow indicates the direction and magnitude of the respective horizontal and vertical velocities

The optical flow methods try to calculate the motion between two image frames which are taken at times t and $t + \Delta t$ at every pixel position. These methods are called differential since they are based on local Taylor series approximations of the image signal; that is, they use partial derivatives with respect to the spatial and temporal coordinates.

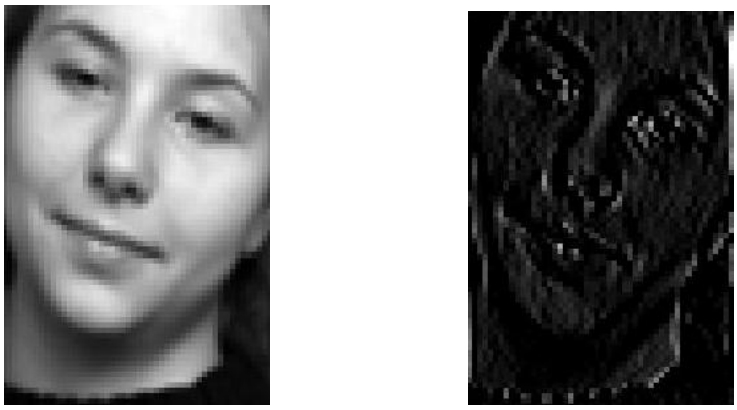
2.3.3 Pre-Whitened image

For removing the cross correlated term in the image sequence, per-whitening is to be done as from the theorem for entropy maximization. Which can be done by multiplying the data by the inverse of the square root of the mean power spectrum in the Fourier domain. Each whiten image is treated as single sample of a dimension equals to no.of pixel. Fisher and Darrel has used pre-whitened images for information theoretic fusion in [1], [3].

2.4 Correspondence Measure of Audio and Video

2.4.1 Pearson Correlation Coefficient

Hershey and Movellan [2] has find per pixel correlation between average acoustic energy of audio, with each pixel in an audio-visual sequence. The have ini-



(a) (b)

Figure 2.2: (a) image from a video sequence. (b) Pre-whitened image here lips eyes nose are accentuated

tially assumed that the density are Gaussian later on they have generalized it and gave the measure for finding correlation between each pixel and energy of audio for a no. of consecutive frames.

$$\rho = \frac{\operatorname{cov}(A,V)}{\sqrt{\operatorname{var}(A)\operatorname{var}(V)}} \tag{2.7}$$

where ρ is the pearson correlation coefficient.

2.4.2 Mutual Information

Mutual dependencies between two RV x and y can be quantifies by estimating mutual information $MI(x,y)$ between the two RV in information theory. Mutual information is defined as.

$$I(x,y) = h(x) + h(y) - h(x,y) \tag{2.8}$$

by manipulating entropy different form of entropy can be seen

$$I(x,y) = h(y) - h(y|x) \tag{2.9}$$

$$I(x,y) = h(x) - h(x|y) \tag{2.10}$$

where $I(x,y)$ is the MI between RV x and y, $h(x)$ is the diffential entropy(also knowns as Kullback-Leibler criterion), $h(y|x)$ is the entropy of RV y condi-

tioned on x, $h(x,y)$ is the joint entropy. Entropy quantifies uncertainty about a given random variable or vector. Mutual information measures the amount of information that one RV conveys to other it quantifies relative uncertainty of one RV with respect to other.

Mutual information is related to Pearson correlation coefficients as

$$MI(X,Y) = -\log(1-\rho(x,y)^2) \tag{2.11}$$

in next chapters more details about Mutual information is given as this is the main criteria for finding correspondence between audio and video in this thesis.

Chapter 3

Association of Signal Level Audio-Video Fusion

Hypothesis Test for Data Association

Statistical Implicit Model for Audio-Visual Fusion

Chapter 3

Association of Signal Level Audio-Video Fusion

3.1 Hypothesis Test for Data Association

A statistical hypothesis test is method for finding statistical inference on basis of observation from the set of RV. This test define a procedure for deciding whether the assumed hypothesis in a region of interest fits or not. Which shows how likely a set of observation to occur if the hypothesis is true.

We have an audio visual sequence, having N frames, each audio and video frames are denoted by A_i and V_i as the $i.i.d.$ samples,where i denotes the discrete time. A_i and V_i are the samples from RV A and V will allow us to cast the hypothesis [13, 14] which is defined as.

$$\begin{aligned} H_0 &: A_i, V_i \sim p(A_i)\, p(V_i) \\ H_1 &: A_i, V_i \sim p(A_i, V_i) \end{aligned} \tag{3.1}$$

where, H_0 stats the joint density of audio and video measurement can ne expressed as product of their marginal density(i.e. they are statistically independent) and H_1 stats that joint density is equivalently associated(i.e. they are statistically dependent). If we have probability density of $p(A_i), p(V_i)$ and $p(A_i, V_i)$ by consistent probability density estimator. Then finding the log-likelihood ration and taking expectation with respect to joint density of A and V gives

$$E\left\{\frac{1}{N}\sum_{i=0}^{N-1}\log\left(\frac{p(A_i,V_i|H_1)}{p(A_i,V_i|H_0)}\right)\right\}$$

$$= E\left\{\frac{1}{N}\sum_{i=0}^{N-1}\log\left(\frac{p(A_i,V_i)}{p(A_i)p(V_i)}\right)\right\} \tag{3.2}$$

$$= I(A;V) \tag{3.3}$$

$$= h(A)+h(V)-h(A,V) \tag{3.4}$$

which, is equivalent to finding mutual information between the RV A and V. Mutual information is expressed in equation(3.4) as a combination of differential entropy $h(A)$, $h(A)$ and $h(A,V)$. Finding mutual information between RV A and V is equivalent to log-likelihood test for given hypothesis. Now, the issue that arise here is that how the joint density between audio and video is model such that it can capture the complex relationship between them.

As in such case when direct density estimation is difficult, because of high dimensional audio and video data and also representation of them are different from each other. Non parametric density estimator [15] are helpful for capturing statistical dependency between them.

3.2 Statistical Implicit Model for Audio-Visual Fusion

Consider a multi-modal scene with back ground interference in each modality, and one joint source which is common in both the observation. The measurement that are acquired from a given observation contain the information from the joint source and the background interference. Here, A and V are the audio and video measurements from the observation $\{X,Y,Z\}$ of multi-modal scene with X and Z as the back ground interference and Y is the common cause [16], [1]. A is associated to the observation of X and Y, V is associated to the observation of Y and Z.

3.2.1 Independent Cause Model

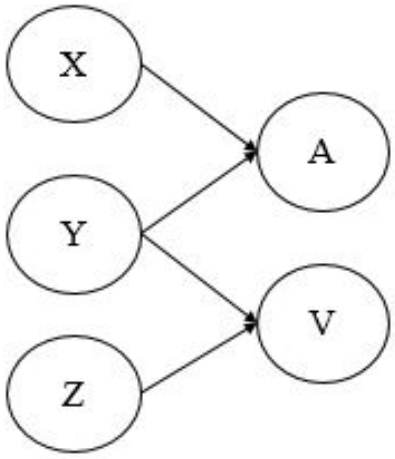


Figure 3.1: Modal shows that A and V are independent but conditioned on $\{X,Y,Z\}$

Fig 3.1, shows the independent cause in which $\{X,Y,Z\}$ are unobserved RV giving the cause for RV $\{A,V\}$. Here, Y is the common cause for both the events. The joint statistical derived from Fig 3.1 is

$$P(X,Y,Z,A,B) = P(X)P(Y)P(Z)P(A|X,Y)P(V|Y,Z) \tag{3.5}$$

3.2.2 Audio is Conditioned on Video

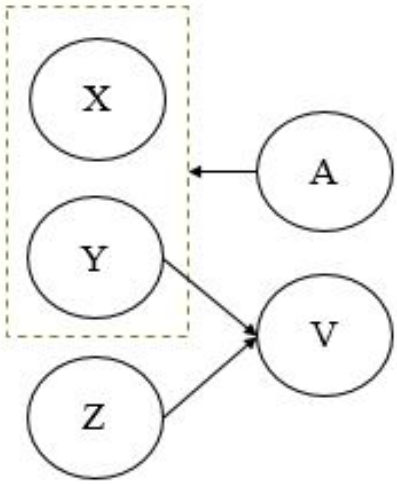


Figure 3.2: information regarding A is given by joint statistics of P and Q

A simple application of bayer's rule in Fig 3.1 yield Fig 3.1, which shows that the information regarding A, is given by the joint statistics of $\{Y,Z\}$ which will be consistent with

$$P(X,Y,Z,A,V) = P(A)P(Z)P(X,Y|A)P(V|Y,Z) \tag{3.6}$$

a similar graph can be obtained if V is conditioned on X and Y.

3.2.3 Resultant Graph by the Existence of Separating Function

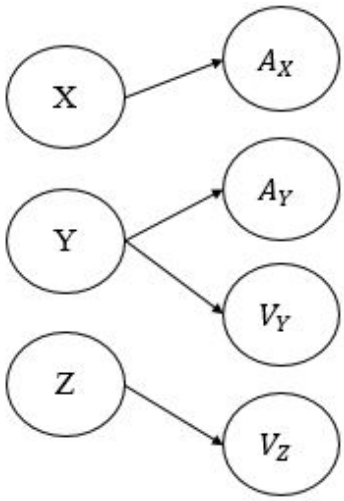


Figure 3.3: resultant graph if separating function exists

If the decomposition of the audio and video measurement exist, then it can see from the Fig 3.3 that there is no influence of X and Z on Y, y is common cause for the same event and is now not influenced by X or Y.

here, $A = [A_X, A_Y]$ and $V = [V_Y, V_Z]$ for this decomposition the joint density can be written as

$$\begin{aligned} P(X,Y,Z,A,V) = P(X)P(Y)P(Z)P(A_X|X) \times \\ P(A_Y|Y)P(V_Y|Y)P(V_Z|Z) \end{aligned} \tag{3.7}$$

3.2.4 Equivalent Markov Chain Model

The required model for which the information theoretic algorithm(equivalent Markov model found by existence of separating function) will be used is given in fig(3.4) . If the decomposition exist than by data processing inequality [17], following inequality holds:

$$\begin{aligned} I(A_Y, V_Y) &\leq I(A_Y, Y) \\ I(A_Y, V_Y) &\leq I(V_Y, Y) \end{aligned} \tag{3.8}$$

also, for any function of A_Y and V_Y (e.g $f_A = f(A_Y, h_a)$, $f_V = f(V_Y, h_V)$) these inequality holds for

$$\begin{aligned} I(f_A, f_V) &\leq I(f_A, Y) \\ I(f_A, f_V) &\leq I(f_V, Y) \end{aligned} \tag{3.9}$$

finally, it can be shown using [13]

$$I(f_A, f_V) \leq I(A_V, V_Y) = I(A, V) \tag{3.10}$$

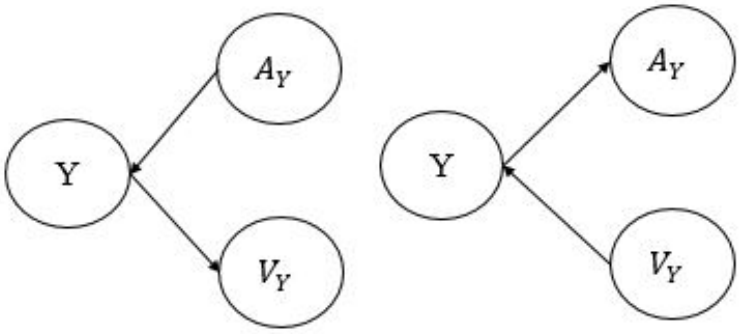


Figure 3.4: equivalent Markov model found by existence of separating function

which, shows that maximizing mutual information between f_V and f_A, $I(f_A, f_V)$ will definitely increase mutual information between F_A and Y, and F_V and Y. These function will be used further for non-parametric estimation of joint density of audio and video in given audio visual sequence.

Chapter 4

Information Theoretic Learning and Feature Extraction

Introduction

Feature Space Representation of Data

Self Organization and Maximizing Entropy

Derivation of ITL algorithm

Method for Solving Under-determined Linear equation

Chapter 4

Information Theoretic Learning and Feature Extraction

4.1 Introduction

The concept of information theoretic learning(ITL) [18], [19], [20] as a criteria for feature extraction is not new. The way of choosing the feature space, leads to useful modification to form iterative learning algorithm. ITL can be used in adaptaton for supervised and unsupervised learning machine. The goal of ITL is to capture as much as information from the given data, in the parameter of learning machine, in order to it to be useful, it should be independent of learning machine and it should not require apriori assumption about data distribution. The approach described[[20] uses Integral Square Error(ISE) and mutual information, as a criterion for adaptation. The method works with any linear or nonlinear mapping which is differentiable in its parameters.

4.2 Feature Space Representation of Data

From the concept of neural network and adaptive filters representation of the feature vectors are done, such that they will used for adaptation criteria to preserve maximum amount of mutual information. It is often that projecting high-dimensional data onto a smaller subspace results in improved performance in a nonparametric classifier.

Let x is the high-dimensional data i.e. $x \in R^N$ is projected in smaller dimen-

sion subspace, by making the function such that $y{=}f(x,h)$, $y \in R^M$ with $M < N$ which is differentiable in h.

The method utilizes Linskers *Principle of InformationMaximization*, which intends to transfer maximum information about a signal from the input to the output of a mapping, as the criterion for feature extraction [21].

4.3 Self Organization and Maximizing Entropy

Linsker's *Principle of InformationMaximization* [21] shows that a mapping of a signal through a neural network should be accomplished so as to preserve the maximum amount of mutual information(MI). The basic requirement for any transformation, through neural network is accomplished in order to maximize the amount of information preserved.

The demand of MI as a criteria of feature extraction is threefold. First, MI takes the advantage of the underlying PDF. Second, adaptation will remove as much as uncertainty about input class by observation of the output $y{=}f(x,h)$. Third, this is achieved with constraint of mapping topologies.
Three equivalent form of MI [17] are

$$I(x,y) = h(x) + h(y) - h(x,y) \tag{4.1}$$

$$I(x,y) = h(y) - h(y|x) \tag{4.2}$$

$$I(x,y) = h(x) - h(x|y) \tag{4.3}$$

where $I(x,y)$ is the MI between RV x and y, $h(x)$ is the diffential entropy(also knowns as Kullback-Leibler criterion), $h(y|x)$ is the entropy of RV y conditioned on x, $h(x,y)$ is the joint entropy. Entropy quantifies uncertainty about a given random variable or vector. Mutual information measures the amount of information that one RV conveys to other it quantifies relative uncertainty of one RV with respect to other.

For a continous RV, $x \in R^N$ entropy is defines as [17]

$$h(u) = -\int_{\infty}^{\infty} \log\left(f_U(u)\right) f_U(u) dx \tag{4.4}$$

where $f_U(u)$ is the probability density function of the RV, and base of logarithm is arbitary. Entropy can also be shown as expected value of log of the probability density function.

$$h(u) = E\left\{\log\left(f_U(u)\right)\right\} \tag{4.5}$$

4.3.1 Properties for Entropy Maximization

1. If the RV is in finite range of $\in R^N$ is restricted, then entropy is maximized for uniform distribution as uniform distribution has the maximum entropy.
2. If in covariance matrix diagonal elements are held constant, then entropy is maximized for normal distribution with diagonal covariance matrix.
In the above given property, the RV should be *Statistically Independent* to each other. First property will be used in this method [24].

4.3.2 Nonparametric PDF Estimation

The difficulty that arise as a criteria of adaptation of mutual information is that it is an integral function of probability i.e. it requires complete knowledge of the PDF for which mutual information is to be estimated. Further more we have not given density itself but from the samples it must be inferred. For which assumption have to made about the form of the density function which is difficult to measure directly.

However if the dimension of RV can be controlled by transforming it to new RV, then Non parametric estimator will give better results as shown in [1], which relies of such estimate in the output space which result in less computational complexity.

Non parametric kernel based estimator such as Parzen window method are

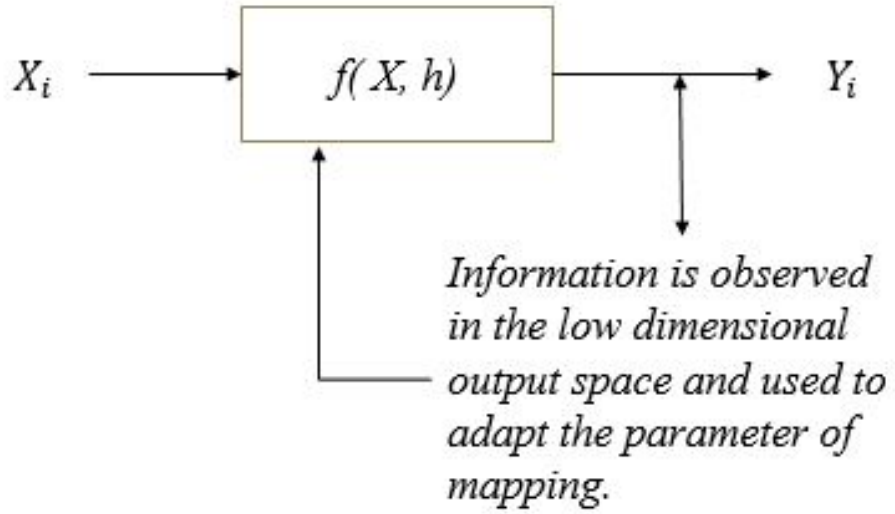


Figure 4.1: Mapping as feature extraction. Information content is measured in the low dimensional space of the observed output.

used for estimating the probability density function which is given as [15].

$$\hat{f}_X(u) = \frac{1}{N_x} \sum_{j=1}^{N_x} k(x_j - u) \tag{4.6}$$

here, $x_i \in \Re^N$ are observation of RV, $k()$ is the guassian kernel which satisfies the property of PDF i.e. $k(u) > 0$ and $\int k(u)\,du = 1$

This kernel estimator can be viewed as the convolution of kernal function with the obsevation. In this method their goal is to local estimate of the PDF i.e kernel should be localized unimodal and decaying to zero, also $k(u)$ should be differentiable every where.

4.3.3 Prior work in Information Theoretic Learning

The concept of information theoretic learning in neural processing as a criteria for feature extraction is not new. Entropy is used as a measure for adaptation. Input RV are mapped in the output subspace and then in the output space entropy is used as measure of adaptation when using neural processing. This

method differ here for neural processing in two ways.
1. The form of density here is discrete while we are working with continuous entropy.
2. This method is using neural processing, in which mapping should be constraint to be linear from input subspace to output subspace, where as the given method can be used with any arbitrary non-linear mapping as long as they are differntiable.

4.4 Derivation of ITL algorithm

Using the property that is show in previous section for the maximization of entropy, that if the RV variable is in finite range then entropy is maximized for uniform density. For adaptation point of view in this method use differential entropy is not worthy, therefore it has to be approximated by its second order Taylor series expansion [1] by utilizing the concept of integrated square error(ISE) between uniform density to the density estimated by Parzen density estimate for entropy maximization. This method is unsupervised in the sense that it require no target distribution, this is an integral square error between uniform density and estimates density. This method is given in [19], [20].

$$\begin{aligned} J &= \tfrac{1}{2}\int_R \left(u(v) - \hat{f}_Y(v,g)\right)^2 dv \\ &= \sum_{j=1}^{L} \left(u(v_j) - \hat{f}_G(v_j,g)\right)^2 \Delta v \end{aligned} \tag{4.7}$$

here, $g = f(x,h)$ is the linear function of x, which is related as

$$g_i = h^T x_i, \forall i \tag{4.8}$$

x is the RV $x \in R^{N_x}$, and $h^T \in R^{N_G \times N_X}$ are the parameters for which the mutual information is maximized that will give low dimension statistics in the output space, $g \in R^{N_G}, N_G < N_X$. $u(v)$ is the uniform probability distribution in the region $R = \left\{v \,\middle|\, \left(g^{\min} \leq v^r \leq g^{\max}, r = 1,..N_G\right)\right\}$, $\left\{g^1, g^2, .., g^{N_G}\right\}$ are the compo-

nents of g. By dividing the entire region containing R in small square domain at $L >> N$ uniformly distributes spatial location $v_j \in R^{N_G}$, j=1,..,L this integral can ne changed into sum. Now, partial derivative of the criteria function by chain rule with respect to h is given as

$$\frac{\partial J}{\partial h} = \left(\frac{\partial J}{\partial \hat{f}}\right)\left(\frac{\partial \hat{f}}{\partial g}\right)\left(\frac{\partial g}{\partial h}\right) \tag{4.9}$$

to minimize the function with respect to mapping parameters. The estimate density for input RV x at a location v by parzen window is given as in [15],

$$\hat{f}_X(v,x) = \left(\frac{1}{N_X}\right)\sum_{i=1}^{N_X} k(x_i - v) \tag{4.10}$$

where, $k()$ is the gaussian kernel, similarly for RV in the output space is given as

$$\hat{f}_G(v,g) = \left(\frac{1}{N_G}\right)\sum_{i=1}^{N_G} k(f(x_i,h) - v) \tag{4.11}$$

non parametric estimate of density in the output space, with linear function of x and h.

$$\frac{\partial J}{\partial h} = -\Delta v\left(\sum_j u(v_j) - \hat{f}_G(v_j,g)\right)\left(\frac{\partial \hat{f}}{\partial g}\right)\left(\frac{\partial g}{\partial h}\right) \tag{4.12}$$

using, equation(4.9) and taking differentiation of criteria function.

$$\frac{\partial \hat{f}}{\partial g} = \left(\frac{1}{N_G}\right)\sum_{i=1}^{N_G} k'(g_i - v_j)$$

$$= \left(\frac{1}{N_G}\right)\sum_{i=1}^{N_G} k'(f(x_i,h) - v_j) \tag{4.13}$$

substituting, equation(4.12) and equation(4.13) in equation(4.9) which comes as.

$$\frac{\partial J}{\partial h} = \left(\frac{\Delta v}{N_G}\right) \sum_j \left(u(v_j) - \hat{f}_G(g, v_j)\right) \sum_i k'(g_i - v_j) \frac{\partial}{\partial h}(f(x_i, h)) \tag{4.14}$$

for minimizing equation(4.14), the measured error ε_i minimizing it, will minimize the criteria function, there error term will give the update term, for which entropy is to be maximized.

$$\varepsilon_i = \sum_j \left(u(v_j) - \hat{f}_G(g, v_j)\right) \frac{\partial}{\partial g_i} \left(\hat{f}_G(g, v_j)\right) \Delta v \tag{4.15}$$

equation(4.8) can also be written as.

$$\frac{\partial J}{\partial h} = \sum_i \frac{\partial J}{\partial g_i} \frac{\partial g_i}{\partial h} \tag{4.16}$$

evaluation of error term also can be possible is the submission about j changes to integral.

$$= -\sum_i \left\{ \int \left[u(v) - \hat{f}_G(g, v)\right] \frac{\partial}{\partial g_i} \hat{f}_G(g, v)\, dy \right\} \frac{\partial}{\partial h} f(x_i, h) \tag{4.17}$$

error term ε_i given as

$$= -\frac{1}{N_G} \sum_i \varepsilon_i \frac{\partial}{\partial h} f(x_i, h) \tag{4.18}$$

which is the convolution difference between uniform density , and parzen density estimate, and derivative of kernel used in parzen density estimation.

$$\varepsilon_i = \varepsilon_G(g, v) * k'(v) \,|\, v = g_i \tag{4.19}$$

$$= \left(f_G(v) - \hat{f}_G(g,v)\right) * k'(v) \,|v = g_i \tag{4.20}$$

$$= \left(f_G(v) - \frac{1}{N_G} g(v) * k(v)\right) * k'(v) \,|v = g_i \tag{4.21}$$

simplifying the expressions,

$$= \left(f_G(v) * k'(v)\right) - \frac{1}{N_G} g(v) * k(v) * k'(v) \,|v = g_i \tag{4.22}$$

$$= f_r(v) - \frac{1}{N_G} g(v) * k_a(v) \,|v = g_i \tag{4.23}$$

$$= f_r(g_i) - \frac{1}{N_G} \sum_{j \neq i} k_a(g_i - g_j) \tag{4.24}$$

$f_r(g_i)$, is the topology regulation function and $k_a()$ is the attractive kernel.

$k_a()$ is the convolution of kernel with its derivative, and it can be easily computed as

$$k_a(v) = k(v) * k'(v) \tag{4.25}$$

$$k_a(v) = -\left(\frac{1}{2^{M+1}\pi^{M/4}\sigma^{(M+2)}}\right) \exp\left(-\frac{1}{4\sigma^2}\left(v^T v\right)\right) v \tag{4.26}$$

$f_r(g_i)$ is convolution of uniform hypercube with derivative of kernel.

$$f_r(v) = f_V(v) * k'(v) \tag{4.27}$$

$$f_r(v) = \int_R f_V(x) k'(v-x)\, dx \tag{4.28}$$

definition of uniform hypercube

$$f_V(v) = \begin{cases} \prod_i d_i - c_i & d_i \leq v_i \leq c_i, \forall i \\ 0 & otherwise \end{cases} \tag{4.29}$$

$$c_i = -d_i = \frac{d}{2} \tag{4.30}$$

using equation(4.28), the following equation can be written as.

$$f_r(v) = \frac{1}{d^N} \int_{-\frac{d}{2}}^{\frac{d}{2}} .... \int_{-\frac{d}{2}}^{\frac{d}{2}} k'(v-x) dx \tag{4.31}$$

$$f_r(v) = \frac{1}{d^N} \int_{-\frac{d}{2}}^{\frac{d}{2}} .... \int_{-\frac{d}{2}}^{\frac{d}{2}} -\frac{1}{(2\pi)^{N/2}\sigma^{N+2}} \exp\left(-\frac{1}{2\sigma^2}(v-x)^T (v-x)\right)(v-x)\, dx \tag{4.32}$$

$$f_r(v) = -\frac{1}{d^N (2\pi)^{N/2}\sigma^{N+2}} \exp\left(-\frac{1}{2\sigma^2}\sum_i (v_i - x_i)^2\right)(v-x)\, dx \tag{4.33}$$

$$f_r(v)_j = -\frac{1}{d^N (2\pi)^{N/2}\sigma^{N+2}} \left(\prod_{i\neq j} \int_{-\frac{d}{2}}^{\frac{d}{2}} \exp\left(-\frac{1}{2\sigma^2}(v_i - x_i)^2\right) dx_i\right) \times ... \tag{4.34}$$

j_{th} element of $f_r(v)$ is given as

$$\int_{-\frac{d}{2}}^{\frac{d}{2}} \exp\left(-\frac{1}{2\sigma^2}(v_j - x_j)^2\right)(v_j - x_j)\,dx_j$$

$$f_r(v)_j = -\frac{1}{d^N(2\pi)^{N/2}\sigma^{N+2}}\left(\prod_{i\neq j}\sigma\sqrt{\frac{\pi}{2}}\left(erf\left(\frac{v_i+\frac{d}{2}}{\sqrt{2}\sigma}\right)\right) - \left(erf\left(\frac{v_i-\frac{d}{2}}{\sqrt{2}\sigma}\right)\right)\right)\times\ldots$$

$$\ldots\sigma^2\left(\exp\left(-\frac{1}{2\sigma^2}\left(v_j-\frac{d}{2}\right)^2\right) - \exp\left(-\frac{1}{2\sigma^2}\left(v_j+\frac{d}{2}\right)^2\right)\right) \tag{4.35}$$

$$f_r(v)_j = -\frac{1}{d^N}\left(\prod_{i\neq j}\frac{1}{2}\left(erf\left(\frac{v_i+\frac{d}{2}}{\sqrt{2}\sigma}\right)\right) - \left(erf\left(\frac{v_i-\frac{d}{2}}{\sqrt{2}\sigma}\right)\right)\right)\times\ldots$$

$$\ldots\left(\frac{1}{\sqrt{2\pi}\sigma}\left(\exp\left(-\frac{1}{2\sigma^2}\left(v_j-\frac{d}{2}\right)^2\right) - \exp\left(-\frac{1}{2\sigma^2}\left(v_j+\frac{d}{2}\right)^2\right)\right)\right) \tag{4.36}$$

$$f_r(v)_j = -\frac{1}{d^N}\left(\prod_{i\neq j}\frac{1}{2}\left(erf\left(\frac{v_i+\frac{d}{2}}{\sqrt{2}\sigma}\right)\right) - \left(erf\left(\frac{v_i-\frac{d}{2}}{\sqrt{2}\sigma}\right)\right)\right)\times$$

$$\left(k_1\left(v_j-\tfrac{d}{2},\sigma\right) - k_1\left(v_j+\tfrac{d}{2},\sigma\right)\right) \tag{4.37}$$

$$f_r(v)_j \approx \frac{1}{d^N}\left(k_1\left(v_j-\frac{d}{2},\sigma\right)_j - k_1\left(v_j+\frac{d}{2},\sigma\right)_j\right) \tag{4.38}$$

$$f_r(v) \approx \begin{bmatrix} \frac{1}{d^N}\left(k_1\left(v_1-\frac{d}{2},\sigma\right)_1 - k_1\left(v_1+\frac{d}{2},\sigma\right)_1\right) \\ \frac{1}{d^N}\left(k_1\left(v_2-\frac{d}{2},\sigma\right)_2 - k_1\left(v_2+\frac{d}{2},\sigma\right)_2\right) \\ . \\ . \\ \frac{1}{d^N}\left(k_1\left(v_N-\frac{d}{2},\sigma\right)_N - k_1\left(v_N+\frac{d}{2},\sigma\right)_N\right) \end{bmatrix} \tag{4.39}$$

and finally the update term derived from ITL is

$$\Delta g_i = f_r\left(g_i^{k-1}\right) - \frac{1}{N}\sum_{i\neq j} k_a\left(g_i^{k-1} - g_j^{k-1}, \Sigma\right) \tag{4.40}$$

$$f_r(g_i)_l = \frac{1}{d}\left(k\left(g_i+\frac{d}{2},\sum\right)_l - k\left(g_i+\frac{d}{2},\sum\right)_l\right) \tag{4.41}$$

$$k_a\left(g,\sum\right) = k\left(g,\sum\right) * k'\left(g,\sum\right)$$

$$= -\frac{1}{2^{M+1}\pi^{M/2}\sigma^{M+2}} \exp\left(\frac{g^T g}{4\sigma^2}\right) g \tag{4.42}$$

after the update of all the term in order to maximize entropy, the methods used result in system of under-determined linear equation which has to be solved for finding projection coefficients.

4.5 Method for Solving Under-determined Linear equation

Consider a set of linear equation as,

$$G = XH \tag{4.43}$$

if there are more no. of variable then equation or there are many solution of equations then it is said to be under-determined equations. Which are solved using least square method for which steps are as follows.

In this case, it is common to seek a solution H with minimum norm. That is, it needed to solve the optimization problem

$$\min_{H} \| H \|_2^2 \tag{4.44}$$

such that $G = XH$

minimization can be done using Lagrange Multiplier, defining the Lagrangian :

$$L(H,\mu) = \min_{H} \| H \|_2^2 + \mu^T (G - HX) \tag{4.45}$$

taking derivative of equation(4.45),

$$\begin{aligned} \frac{\partial L(H)}{\partial H} &= 2H - X^T\mu \\ \frac{\partial L(\mu)}{\partial \mu} &= G - XH \end{aligned} \tag{4.46}$$

setting the derivative to zero,

$$\begin{aligned} H &= \tfrac{1}{2}X^T\mu \\ G &= XH \end{aligned} \tag{4.47}$$

which gives

$$G = \frac{1}{2}XX^T\mu \tag{4.48}$$

then

$$\mu = 2\left(XX^T\right)^{-1}G \tag{4.49}$$

Finally

$$H = X^T \left(XX^T\right)^{-1} G \tag{4.50}$$

this will give the coefficient of projection, after which image of projection coefficient is formed, which gives the region in which maximum projection between audio and video are for correct audio it will concentrated, more on speakers lips and for incorrect audio projection are spreaded on the entire image.

Chapter 5

Joint Processing of Audio and Video by Signal Level Fusion

Introduction

Projection of Data from Input Subspace to Output Subspace

Chapter 5

Joint Processing of Audio and Video by Signal Level Fusion

5.1 Introduction

The objective of signal level fusion is employed to produce a result that provide most reliable and detailed information regarding the data from different domain than getting information from individual domain. Signal level fusion techniques combines data from different domain together. This will give the efficient representation of the given data. In this method derived feature from data are fused.

5.2 Projection of Data from Input Subspace to Output Subspace

In the previous chapter we have discuss the theorem for entropy maximization, that if a RV is in finite range then entropy is maximized for uniform distribution. The main difficulty arise here is that, audio and video representation have different dimensions and size. One thing that is common in both measurement is that they have sampled at same frame rate($29.97 fps$). Dimension of a video in particular frame is equal to the no. of pixel in that frame, and of audio is dependent to audio sampling frequency, if a audio is samples at $16Khz$ as in given in CUAVE audio visual dataset, then in particular audio frame there are 533 samples.

5.2.1 Subspace Projection

Each video frame is converted into single vector of dimension equal to no. of pixels and i_{th} video frame is denoted by V_i. similarly audio is converted into sequence of periodogram(magnitude to windowed FFTs) taken at video frame rate and whose i_{th} audio frame is denoted by A_i. For entropy maximization RV should be $StatisticallyIndependent$ owing to which prewhitening transformation has to be applied on video measurement(method given in chapter 2), for removing the cross correlated terms then after it can be used for this method.

The approach given in[[3], [1], projection for audio/video is given as

$$f_{V_i} = h_V^T V_i \tag{5.1}$$

$$f_{A_i} = h_A^T A_i \tag{5.2}$$

here, $V_i \in \Re^{N_V}$ and $A_i \in \Re^{N_A}$ are samples of image and audio periodogram from an audiovisual sequence. f_{V_i} and f_{A_i} are resultant low-dimensional feature, whose dimensionality is determined by projection matrix h_V^T and h_V^T. $h_V^T = \Re^{M_v \times N_V}$ and $h_V^T = \Re^{M_v \times N_V}$ results in $f_{V_i} = \Re^{M_V}$, $f_{A_i} = \Re^{M_A}$ dimension feature vector in output space with $M_V < N_V$ and $M_A < N_A$.

Such transformation will allow to use the method described in previous chapter under all constraint. Now, treating V_i and A_i as a sample from RV V and A, our goal is to estimate h_V and h_A such that it will maximize the mutual information between V and A.

All the updation of the feature vector are done in the output space, the method require the transformation will have the unique inverse, in order to calculate the value's of h_V and h_A. The results of such projection can be shown by forming the images of h_V by which it can be shown that both data belongs to the same event or not.

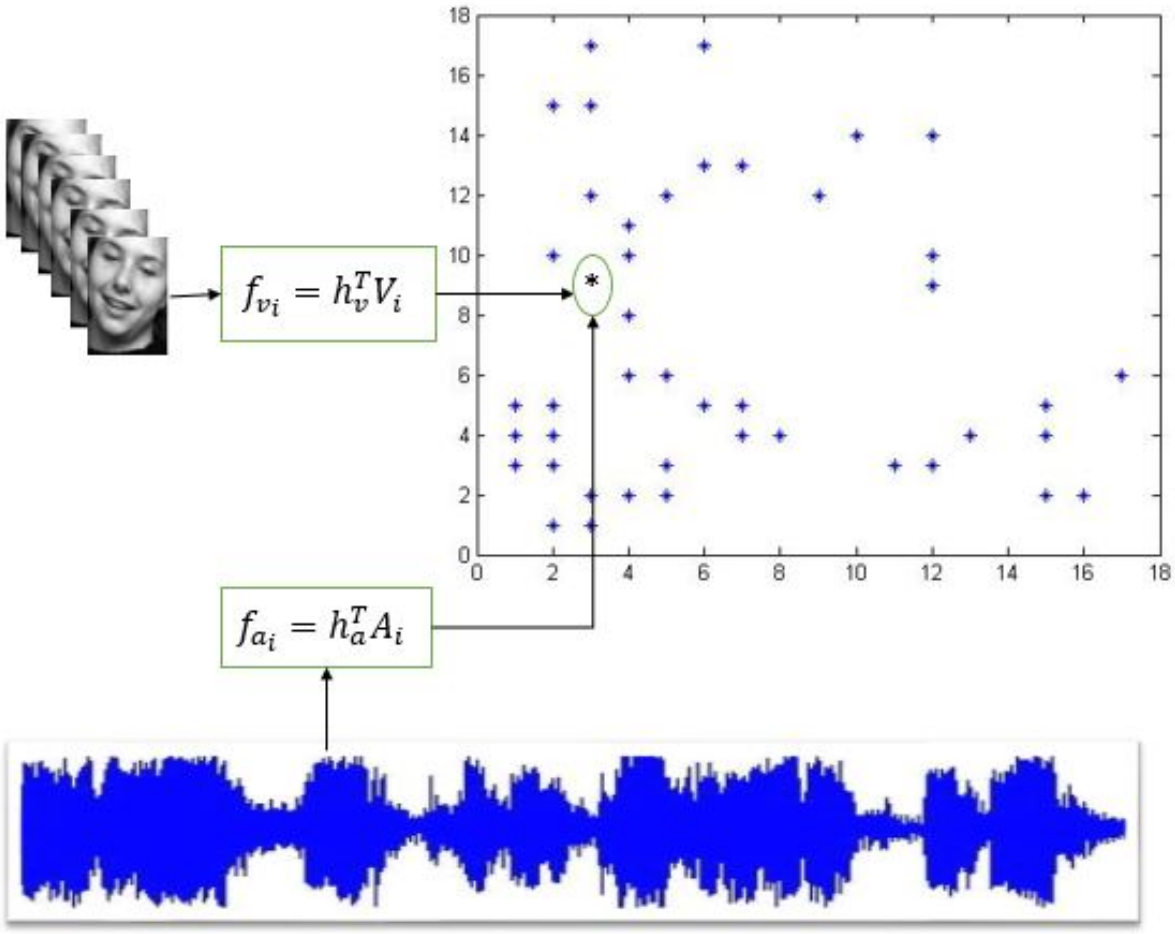


Figure 5.1: shows the projection of the audio and video frame to low dimensional output space, in which learning has to be done

5.3 Information Theoretic Learning

The measure of information that conveyed from one RV to another can be found by calculating the mutual information between them. Till here transforming high dimensional RV to low dimensional subspace is done. Now, it is required to find the mutual dependencies between a pair of feature vector in low dimensional subspace. Which can be found by calculating mutual information between a pair of feature vector.

In previous chapter different way of calculating mutual information as a

combination of different form of entropy is shown.

5.3.1 Mutual Information definition

Mutual information as a combination of differential entropy is expressed as [17]

$$I(f^v, f^a) = h(f^v) + h(f^a) - h(f^v, f^a) \tag{5.3}$$

when the probability density function of a RV is given this equation cab be used.

$$\begin{aligned} I(f^v, f^a) = & \int_{R_{fa}} Pf^a(x) \log(Pf^a(x)) dx \\ & + \int_{R_{fv}} Pf^v(x) \log(Pf^v(x)) dx \\ & - \iint_{R_{fa} \times R_{fv}} p_{f^a, f^v}(x,y) \log\left(p_{f^a, f^v}(x,y)\right) dxdy \end{aligned} \tag{5.4}$$

For adaptation point of view equation(5.4) cannot be used, because it is function of probability density. It requires knowledge about the form of density, whose estimation is difficult as in this case unless assumption are made. For this, in this method each entropy is replaced by its second order taylor series approximation [1].

$$\widehat{I}(f^a, f^v) = \widehat{H}(f^a) + \widehat{H}(f^v) - \widehat{H}(f^a, f^v) \tag{5.5}$$

second order taylor series approximation of mutual information as a intergral difference between uniform density to the Parzen density estimate is given by

$$\begin{aligned}\widehat{I}(f^a, f^v) = &\int_{R_{fa}} \left(\widehat{p}_{f^a}(x) - p_u(x)\right)^2 dx \\ &+ \int_{R_{fv}} \left(\widehat{p}_{f^v}(x) - p_u(x)\right)^2 dx \\ &- \int_{R_{fa}\times R_{fv}} \left(\widehat{p}_{f^a,f^v}(x,y) - p_u(x,y)\right)^2 dxdy\end{aligned} \tag{5.6}$$

The approximation used for individual entropy term is integral square campare between uniform density to the density of projection. The gradient term of this approximation with respect to projection coefficient can be calculated by a finite no. of function at a finite no. of location in output space. The derivation of this update value in given in previous chapter for each feature vector as k_{th} iteration for i_{th} feature vector as a combination of feature vactor value at $k-1$ iteration is given as

$$\Delta f_i = f_r\left(f_i^{k-1}\right) - \frac{1}{N}\sum_{i\neq j} k_a\left(f_i^{k-1} - f_j^{k-1}, \sum\right) \tag{5.7}$$

$$f_r(f_i)_l = \frac{1}{d}\left(k\left(f_i + \frac{d}{2}, \sum\right)_l - k\left(f_i + \frac{d}{2}, \sum\right)_l\right) \tag{5.8}$$

$f_r(f_i)_l$ is the element of l_{th} of $f_r(f_i)$ and it is M dimensional vector value function, M depends on the projection matrix.

$$\begin{aligned}k_a\left(f, \sum\right) &= k\left(f, \sum\right) * k'\left(f, \sum\right) \\ &= -\frac{1}{2^{M+1}\pi^{M/2}\sigma^{M+2}} \exp\left(\frac{f^T f}{4\sigma^2}\right) f\end{aligned} \tag{5.9}$$

$k_a(f_i, \sigma)$ is the attractor kernel, and it is the convolutiom of kernel with its

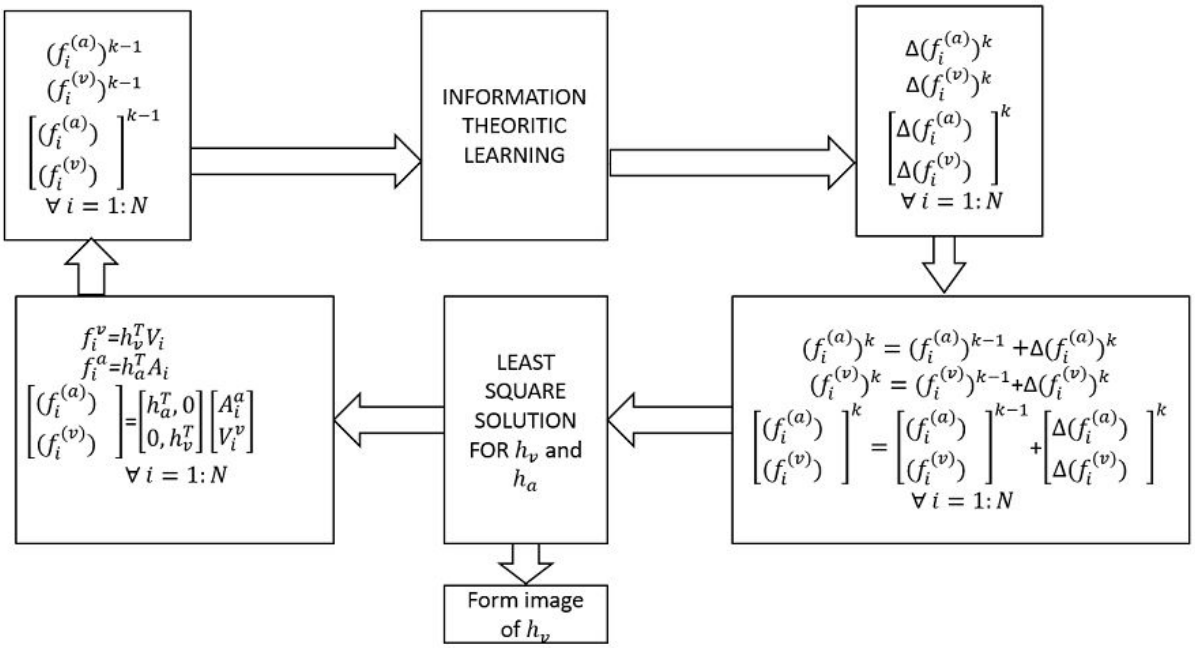


Figure 5.2: shows the learning procedure and calculation of the projection coefficient adaptively

derivative, it is also M dimensional vector value function, d is the uniform PDF in the output space. Here, value of M_v and M_a are set to 1, i.e. the dimension of f_v and f_a are set to unity. Experiment follow the update rule above then finding the value of h_v and h_a by least square method, actually this method results in system of underdetermined equation, there are more no. of variable than no. of equations, which can be estimated by least square method.

The magnitude value of these projection coefficient will show the relation between the audio and video being fused. If video is fused with correct audio then projection will be concentrated more on speaker lips means the magnitude value of projection are more across speaker lips where as if incorrect or alternate audio is used these projection are spreaded on the entire region of the image of h_v which will give the consistency og audio and video being fused.

Experimental results are show in the next chapter.

Chapter 6

Simulation Results

Result for speaker verification using CUAVE audio visual dataset

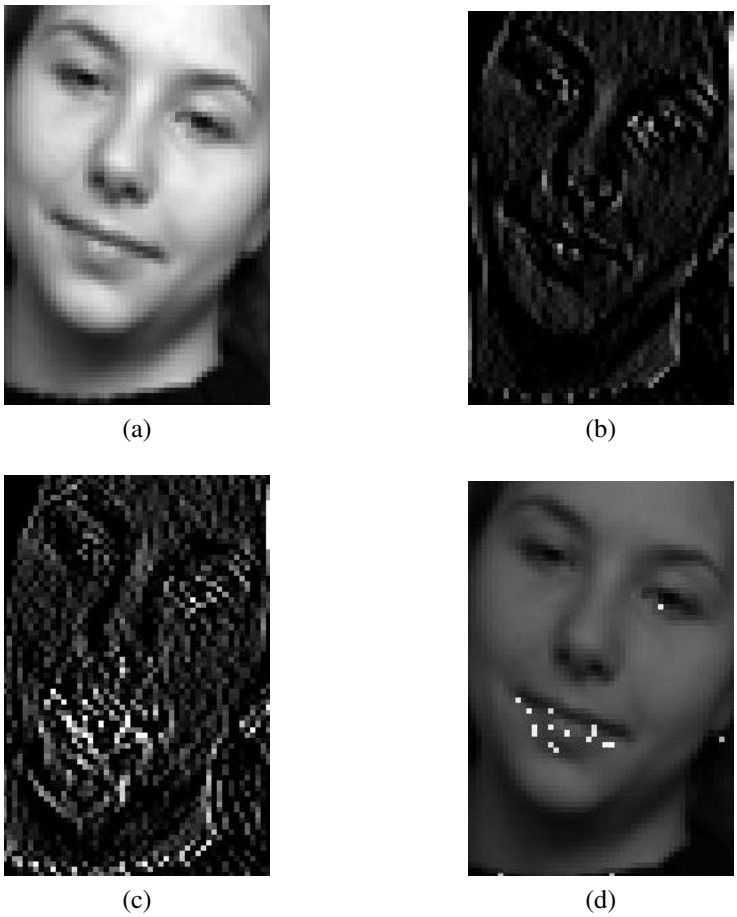
Chapter 6
Simulation Results

We have done experiment using CUAVE database. In this database video data was collected at 29.97 frames per second with 75 $\times$ 50 resolution. The audio signal was collected at $16KHz$. In this video data is consisting of two speakers where the audio belongs to one of them. Firstly, we separate the video into two segments where the audio belongs to the first video. Secondly, the audio signal was transformed to a series of periodograms. The window length of periodogram was 2/29.97s.

After Maximization of the mutual information between the projected audio and video data samples, projection coefficients are used as a measure of consistency. Results are consist in four step, first image in any figure is an image from the datset in which operation is to be performed, second image is the pre-whitened image for removing cross correlated term in image and it is applied to all the image, third is the image of the projection coefficient which shows the region for which there are maximum projection values through which verification is done, fourth image is the binary threshold image which shows the region which have maximum projections.

This is a two way experiment in each data set there are two video sequence and one audio sequence, results show what happens when correct audio and video is fused and also when in correct audio is used for a video sequence. For different video sequences experiments are carried out.

`http://people.csail.mit.edu/siracusa/avdata/`



(a) (b)

(c) (d)

Figure 6.1: (a) Image from video sequence. (b) Prewhitened image. (c) Image of projections when fused with the correct audio. (d) After binary thresholding of projection mapped to image with threshold 0.95.

Fig. 6.1(a) shows a single frame from first segment of video sequence. Fig. 6.1(b) shows prewhitened image of the frame. We can see the moving edges i.e. lips, chin, etc... are accentuated. Fig 6.1(c) shows image of projections when fused with the audio. As can be seen projection highlights the region about speakers lips. Fig 6.1(d) shows the image with regions which have maximum projection are highlighted. For correct audio we can see maximum projection around speaker lips.

(a) (b)

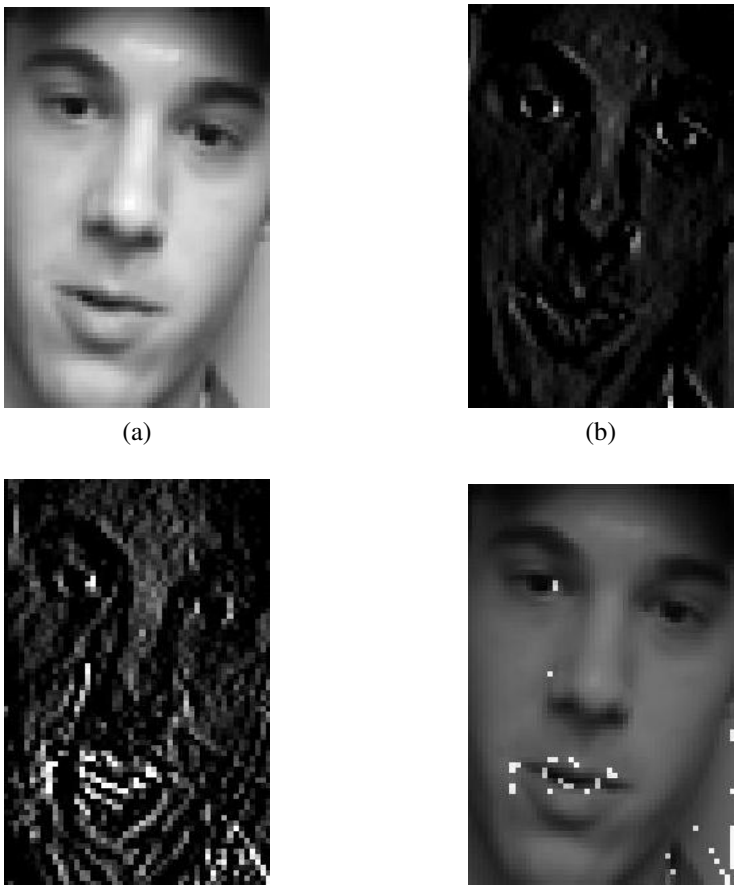
(c) (d)

(e)

Figure 6.2: (a) Image from video sequence. (b) Prewhitened image. (c) Image of projections when fused with the alternate audio. (d) After binary thresholding of projection mapped to image with threshod 0.95. (e) audio of this video is fused

Fig. 6.2(a) shows a single frame from second segment of video sequence. Fig. 6.2(b) shows prewhitened image of the frame. We can see the moving edges i.e. lips, chin, etc... are accentuated. Fig 6.2(c) shows image of projections when fused with the audio. As can be seen projection highlights the random regions. Fig 6.2(d) shows the image with regions which have maximum projection are highlighted.

Similar result are carried out in all the coming audio-visual data set, description of result are summarized in the Figures.



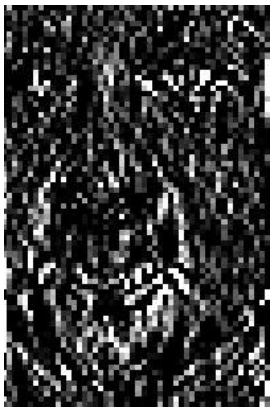
(a) (b)

(c) (d)

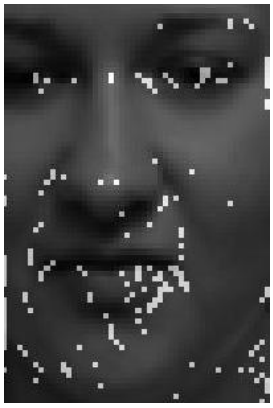
Figure 6.3: (a) Image from video sequence. (b) Prewhitened image. (c) Image of projections when fused with the correct audio. (d) After binary thresholding of projection mapped to image.

(a)

(b)



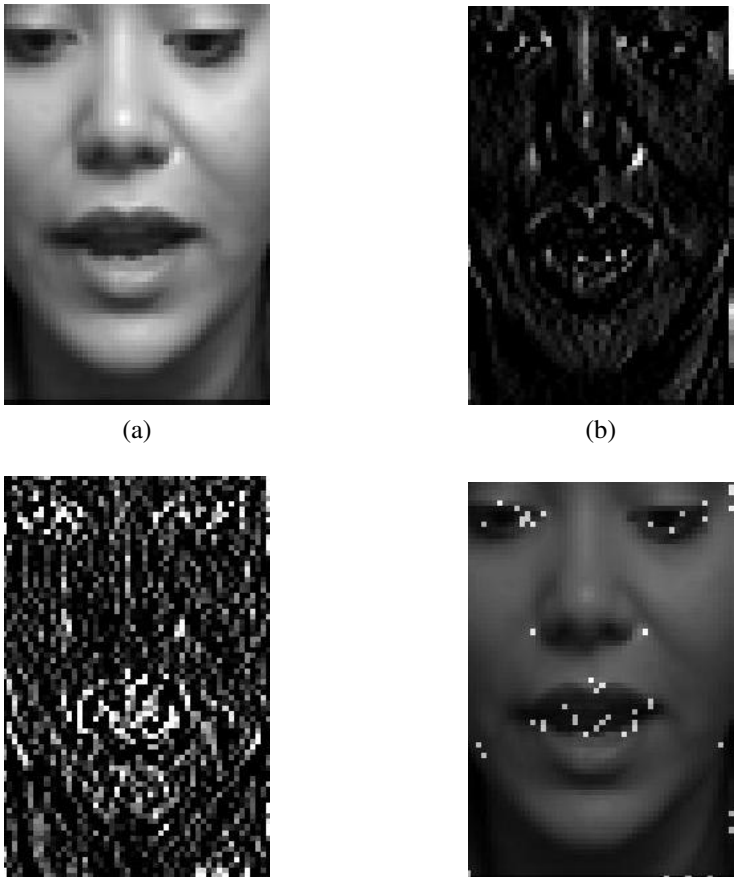
(c)



(d)

(e)

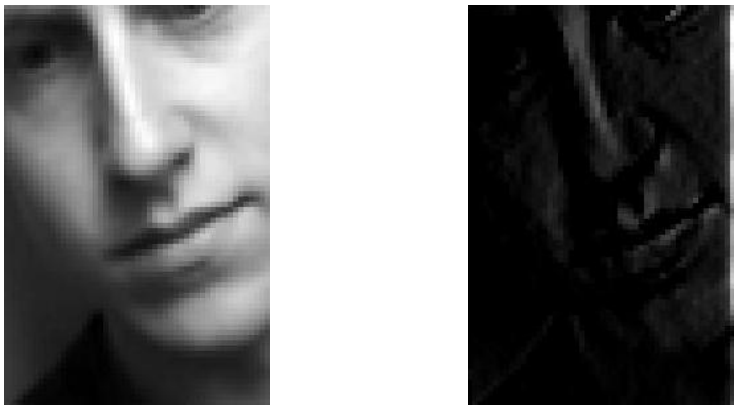
Figure 6.4: (a) Image from video sequence. (b) Prewhitened image. (c) Image of projections when fused with alternate audio. (d) After binary thresholding of projection mapped to image. (e) audio of this video is fused.



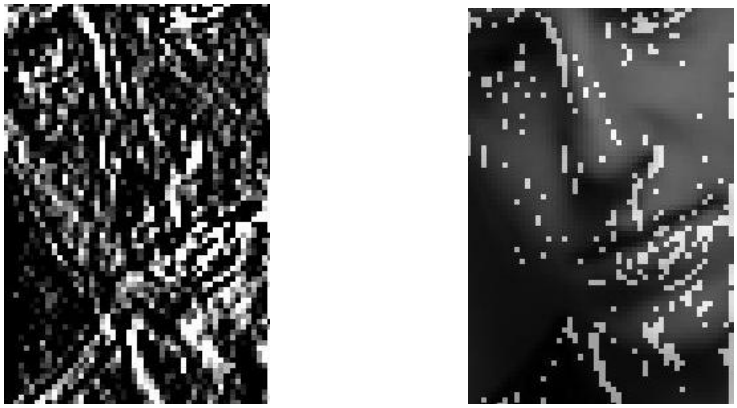
(a) (b)

(c) (d)

Figure 6.5: (a) Image from video sequence. (b) Prewhitened image. (c) Image of projections when fused with the correct audio. (d) After binary thresholding of projection mapped to image.



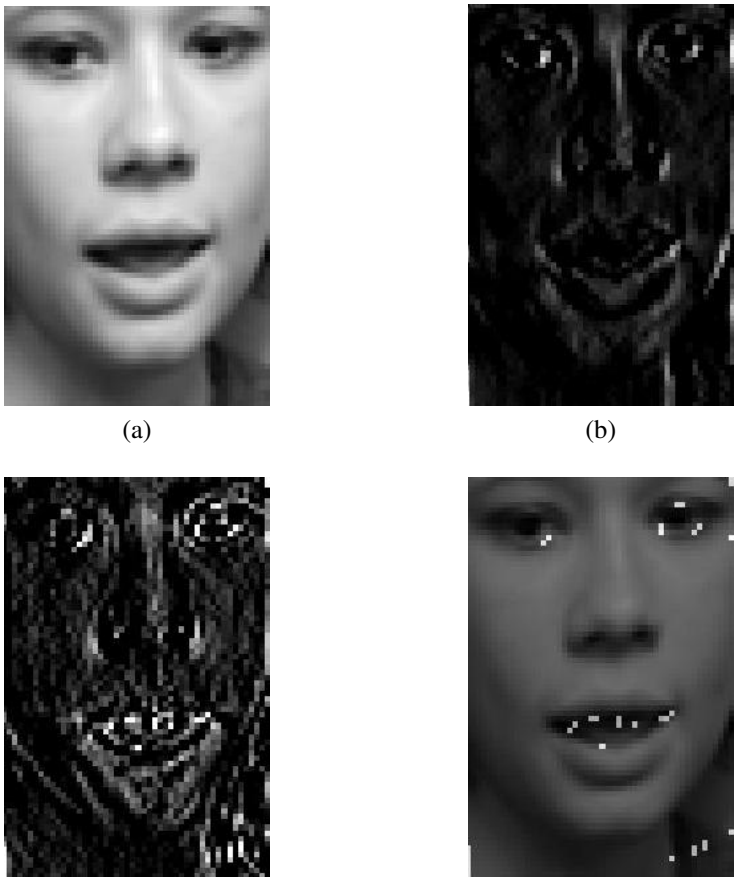
(a) (b)



(c) (d)

(e)

Figure 6.6: (a) Image from video sequence. (b) Prewhitened image. (c) Image of projections when fused with alternate audio. (d) After binary thresholding of projection mapped to image.(e) audio of this video is fused.

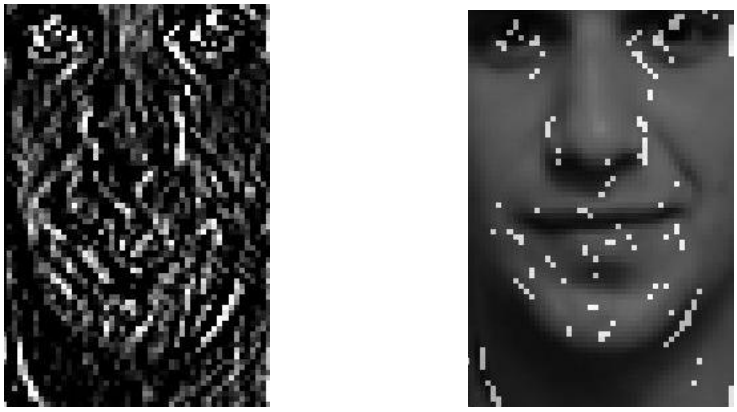


(a) (b)

(c) (d)

Figure 6.7: (a) Image from video sequence. (b) Prewhitened image. (c) Image of projections when fused with the correct audio. (d) After binary thresholding of projection mapped to image.

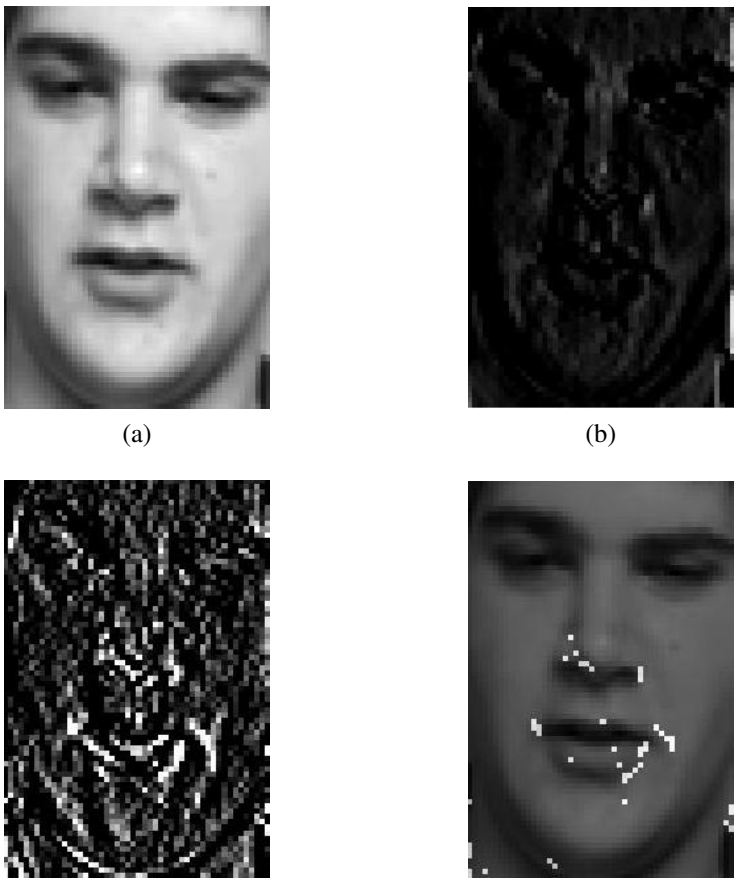
(a) (b)



(c) (d)

(e)

Figure 6.8: (a) Image from video sequence. (b) Prewhitened image. (c) Image of projections when fused with alternate audio. (d) After binary thresholding of projection mapped to image.(e) audio of this video is fused.



(a) (b)

(c) (d)

Figure 6.9: (a) Image from video sequence. (b) Prewhitened image. (c) Image of projections when fused with the correct audio. (d) After binary thresholding of projection mapped to image.

(a)

(b)

(c)

(d)

(e)

Figure 6.10: (a) Image from video sequence. (b) Prewhitened image. (c) Image of projections when fused with alternate audio. (d) After binary thresholding of projection mapped to image. (e) audio of this video is fused.

Chapter 7

Conclusion and Future work

Conclusion

Future work

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this work a technique for measuring the signal level correspondence between audio and video observations is applied on a video from CUAVE database. Experimental results demonstrated that the above method can be used to determine whether a separately recorded video and audio fragments come from the same speaker or not. In this approach signal level fusion is made. There is no restriction of using any acoustic or visual model .i.e a complete video and audio frame is used without any filtering or segmenting video for lip tracking and is not language specific. The domains where such assumptions are viable, or when prior models of individual users are available, such information could be used suitably here. This method is applicable if there is continues audio and video signal for a duration of time i.e. no monologues should be there in between with little head or body movement. This method promise for audio-video consistency and also able to verify that both the audio and video correspond to the same event or not. This method does not make any strong assumption about the underlying joint properties of the modalities being fused(e.g. Gaussian statistics). Here adaptation occurs for a short period of time(approx 2-2.5 sec) of audio-visual data. This method doesn't require any specific training for recognition, all of which it need is a continuous audio-video data for a duration of time.

7.2 Future work

Future work will address the robustness of method over a larger corpus of data. As this method is applicable for only small head and body movement, and no training is required, this method can be extended toward natural and untethered interfaces, where multiple user can interact with causal conversation without attachment or explicit segmentation clues.

Bibliography

[1] J. W. Fisher and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Transactions on Multimedia,*, vol. 6, no. 3, pp. 406–413, 2004.

[2] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," in *Advances in Neural Information Processing Systems 12*, Citeseer, 2000.

[3] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. A. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *NIPS*, pp. 772–778, 2000.

[4] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," in *NIPS*, pp. 814–820, 2000.

[5] H. J. Nock, G. Iyengar, and C. Neti, "Assessing face and speech consistency for monologue detection in video," in *Proceedings of the tenth ACM international conference on Multimedia*, pp. 303–306, ACM, 2002.

[6] J. P. Barker and F. Berthommier, "Evidence of correlation between acoustic and visual features of speech," *Ohala et al*, pp. 199–202, 1999.

[7] H. Bredin, A. Miguel, I. H. Witten, and G. Chollet, "Detecting replay attacks in audiovisual identity verification," in *2006 IEEE International Con-*

ference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings., vol. 1, pp. I–I, IEEE, 2006.

[8] H. Izadinia, I. Saleemi, and M. Shah, "Multimodal analysis for identification and segmentation of moving-sounding objects," *IEEE Transactions on Multimedia,*, vol. 15, no. 2, pp. 378–390, 2013.

[9] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *arXiv preprint arXiv:1003.4083*, 2010.

[10] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[11] E. Gopi, *Digital Speech Processing Using Matlab*. Springer, 2014.

[12] B. K. Horn and B. G. Schunck, "Determining optical flow," in *1981 Technical Symposium East*, pp. 319–331, International Society for Optics and Photonics, 1981.

[13] A. T. Ihler, J. W. Fisher, and A. S. Willsky, "Hypothesis testing over factorizations for data association," in *Information Processing in Sensor Networks*, pp. 239–253, Springer, 2003.

[14] A. T. Ihler, J. W. Fisher, and A. S. Willsky, "Nonparametric hypothesis tests for statistical dependency," *IEEE Transactions on Signal Processing,*, vol. 52, no. 8, pp. 2234–2249, 2004.

[15] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, pp. 1065–1076, 1962.

[16] J. W. Fisher III and T. Darrell, "Probabalistic models and informative subspaces for audiovisual correspondence," in *Computer VisionECCV 2002*, pp. 592–603, Springer, 2002.

[17] T. M and J. A. C. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[18] J. C. Principe, D. Xu, and J. Fisher, "Information theoretic learning," *Unsupervised adaptive filtering*, vol. 1, pp. 265–319, 2000.

[19] J. W. Fisher III and J. C. Principe, "Unsupervised learning for nonlinear synthetic discriminant functions," in *Aerospace/Defense Sensing and Controls*, pp. 2–13, International Society for Optics and Photonics, 1996.

[20] J. W. Fisher and J. C. Principe, "A methodology for information theoretic feature extraction," in *The 1998 IEEE International Joint Conference on Neural Networks Proceedings, 1998.IEEE World Congress on Computational Intelligence.*, vol. 3, pp. 1712–1716, IEEE, 1998.

[21] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, no. 3, pp. 105–117, 1988.

[22] J. W. Fisher and J. C. Principe, "Entropy manipulation of arbitrary nonlinear mappings," in *Proceedings of the 1997 IEEE Workshop Neural Networks for Signal Processing [1997] VII.*, pp. 14–23, IEEE, 1997.

[23] S. Kullback, *Information theory and statistics*. Courier Corporation, 1997.

[24] M. D. Plumbley, *On information theory and unsupervised neural networks*. University of Cambridge, Department of Engineering, 1991.

[25] R. Linsker, "How to generate ordered maps by maximizing the mutual information between input and output signals," *Neural computation*, vol. 1, no. 3, pp. 402–411, 1989.

[26] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling.," in *ISMIR*, 2000.